Design and Implementation of Chinese Historical Text Mining System Based on Culturomics

Lin Tang[1,2](✉) and Chonghui Guo[1]

[1] Institute of Systems Engineering, Dalian University of Technology, Dalian, China
{tanglin,dlutguo}@dlut.edu.cn
[2] Department of Software Engineer, Dalian University of Technology City Institute,
Dalian, China

Abstract. Culturomics and Chinese text mining methods are of great significance for analyzing the development and evolution of Chinese history and culture. To help researchers analyze a large number of Chinese historical text data, a Chinese historical text mining system based on cultruomics is designed, which includes text data processing and analyzing subsystem, text data visualizing subsystem, and text data clustering and retrieval subsystem. First of all, our system preprocesses the text data, then visualizes the text data with the frequency of words line chart and word cloud, at last selects the text data through clustering and retrieval methods. It further supports researchers to discover knowledge from a large number of historical text data. We demonstrate its general performance on text data of Canton Customs into our system. The result shows that our system is feasible and effective.

Keywords: Culturomics · Canton customs · Text mining

1 Introduction

With a long history, China remains in a variety of forms to preserve its culture, such as textural and material researches. In the current research of Chinese culture, the traditional methods are still widely used. Most of them are used to make an empirical summary and inferential exploration through reading and synthesizing a few documents. But to the great extent, they are limited by researchers' subjective understanding and thinking [1]. Because of the vast number of related materials, it is impossible to read them all, which hinders researchers to fully study and interpret history.

Recently, with the rapid development of computer and information technology, researchers begin to use it to investigate society and behavior through social simulation, social network analysis and social media analysis. It gradually forms an interdisciplinary science named computational social science(CSS) [2]. In the culture field of CSS, a new branch cluturomics [3] has appeared. Culturomics is a compound word that consists of culture and genomics, and it is built in mathematics method through the quantitative analysis of digitized texts to study human behavior and cultural trends.

Erez Lieberman Aiden and Jean-Baptiste Michel, two members of the evolutionary dynamics team, made a great contribution in the process of the development of

W. Che et al. (Eds.): ICYCSEE 2016, Part I, CCIS 623, pp. 314–325, 2016.
DOI: 10.1007/978-981-10-2053-7_28

cultruomics. At first Aiden researched genomics through mathematics, then used mathematical tools in evolutionary biology to research historical culture through the quantitative analysis. Moreover, in 2005 Google Book Library, team members collected 5, 195, 769 books, accounting for about 4 % of the total number of books published between 1500 and 2008 in the world [4]. This project established a huge database and offered a visualizer named Google Ngram Viewer(abbreviated as N-Gram) [5] which is one of the most important cultruomics tools. It can query by a term which comprises one or more words. The results are the frequency of the query terms per year. Finally, the frequency is visualized by a continuous time dimension. Michel et al., applied N-Gram into the analysis of culture.

It has got a certain achievement to do the research of culture through the quantitative approach. Relatively, there is scarcity of Chinese historical articles in Google database. N-Gram's concept derives from English grammar, which differs a lot from Chinese grammar. Hence, this tool is inappropriate to direct quantitative analysis of the Chinese articles.

The integrated process of traditional Chinese text mining includes texts preprocessing, indexing and storage, intermediate representation, post-processing [6], which is mainly applied in auto-index, auto-abstract, information retrieval, information extract, document organizing, theme tracking, etc. [7]. Currently, some Chinese text data mining system products and prototype system have emerged, for example, PULSE developed by Gamon and his team in Microsoft [8], mining system of Chinese software commentary developed by Wen Tao and the team [9], etc. But the mining system products or prototype system, which is developed by the historical and culture research based on Chinese text data, has not brought much attention to the public.

We design and implement a Chinese historical text mining system that is based on Culturomics concept can efficiently help researchers to study cultural trends. Our system is designed on the basis of the entire cycle of data mining, including the data collection, text cleaning, text analysis and visualization. The corpus is Chinese historical text with time tags, which will provide multiple visual results for research. Compared with the existing Culturomics tool, our system has the following advantages:

- Our system lays the foundation on Chinese grammar, dealing with the Chinese text in a more efficient way.
- It provides abundant interaction approaches and multiple visualization results to present the results of quantitative data analysis in different aspects.
- It communicates with users to understand their intentions, and selects out what users are interested in through its partitioned clustering and search functions.

2 Design of Chinese Historical Text Mining System Based on Culturomics

According to different research requirements from various clients, including domain specialists and archivists, Chinese historical test mining system consists of three components: Text data processing and analyzing; Text data visualizing; Text data clustering and retrievalling. Our system architecture inspired by CWMS [10], a mining cloud service platform, was constructed in 2013 by Institute of Computer Science, Chinese Academy of Science, but it has some customization. The system includes three subsystems.

1. Text process and analysis subsystem. According to the word frequency statistics, it quantifies the text as the fundamentals for researchers. It has models of Chinese words segmentation, statistic of word frequency and modeling, feature extraction, stop word list maintenance, equal synonyms maintenance.
2. Text visualization subsystem uses word cloud and line chart to make the text visible, so that researchers can easily discover the pattern. It has models of word cloud, line chart of word frequency based on time series.
3. Text clustering and text retrieval system, with the function of text clustering and Full-text retrieval, helps the users find out the text they are interested in.

Figure 1 shows the structure of Chinese text mining system based on Culturomics.

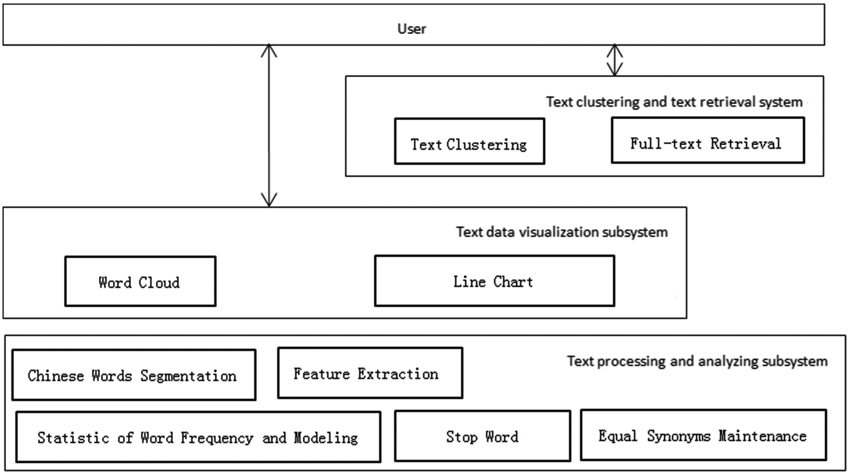
User
Text clustering and text retrieval system
Text Clustering
Full-text Retrieval
Text data visualization subsystem
Word Cloud
Line Chart
Chinese Words Segmentation
Feature Extraction
Text processing and analyzing subsystem
Statistic of Word Frequency and Modeling
Stop Word
Equal Synonyms Maintenance

Fig. 1. Design of chinese historical text mining system based on culturomics

Text data processing section uses word segment system ICTCLAS [11]. Based on cascaded hidden markov model, ICTCLAS is developed by Dr Huaping Zhang and Dr Qun Liu from Institute of Computer technology, Chinese Academy of Science. It plays the leading part in the related domain. Data analysis uses TF-IDF(Term Frequency-Inverse Document Frequency) and normalization to process the data. The output is the quantitative analysis result with time tags. The result is presented through information visualization method such as line chart and word cloud. Users obtain their interested information as character by interacting with the system. Our system uses Lucene(a high-performance text, full-featured text search engine library) [12]. Lucene and K-means clustering algorithm is useful to filter their text.

2.1 Text Data Processing and Analyzing Subsystem

Based on text data processing and analyzing subsystem, our system is used for preprocessing Chinese text. Chinese words segmentation is executed while text is added to our system, then the results are processed in the function of word frequency statistics and modeling.

Words segmentation can break up the input text into the minimal recognizable meaningful unit through computer method. To segment English text is very simple, because of the whitespace between words. But it is difficult for Chinese text because of the lack of whitespace. As the foundation of text mining, the accuracy of Words segmentation directly determines the effect of text mining. Therefore, our system adopts the tokenizer ICTCLAS.

Word set with part-of-speech tagging is the result of the words segmentation. In the process of Words Segmentation, our system deletes the stop words list, combines the synonymous list and extracts Nouns among them.

The most commonly used algorithms [13] in text semantic analysis are Latent Dirichlet Allocation(LDA) and TF-IDF. But TF-IDF is simpler and more efficient than LDA. Thus, We chose the TF-IDF algorithm [14] to accurate the term weighting. The term weighting in each document is:

$$w_{ik} = tf_{ik} \cdot idf_{ik} = \frac{tf_{ik} \cdot \lg\left(\frac{N_i}{n_k}\right)}{\sqrt{\sum_{j=1}^{m}\left(tf_{ij} \cdot \lg\left(\frac{N_i}{n_j}\right)\right)^2}}, i = 1, 2, \ldots, k = 1, 2, \ldots \quad (1)$$

Where the element tf_{ik} is the term frequency of t_k in the document d_i. The element idf_{ik} is the inverse documentation frequency of t_k. N_i is the total number of terms in the I document(the repeated terms still be count). n_k is the frequency of the computed term k.

2.2 Text Visualization Subsystem

The textual data based on a weighted metric can be visualized by different methods. The word cloud [15] can depict the representative keywords at a time. The line chart can illustrate the terms evolution over time. In the historical development, Knowledge is implied not only inside a document, but also in a document set. Therefore, our system combines word cloud and line chart through interacting with users.

The huge differences among texts lengths lead us to normalize the words frequency before the visualization.

$$tf_k^{new} = \frac{tf_k}{\sum_{j=0}^{m} tf_j}, \mathrm{k} = 1, 2, \ldots \quad (2)$$

In the equation tf_k is the term frequency of k. The total number of frequency is $\sum_{j=0}^{m} tf_j$.

We implement the word cloud with d3.js (or just D3 for Data-Driven Documents) [16], which is a JavaScript library for producing dynamic, interactive data visualizations in web browsers. Our word cloud can visualize all terms(selected the top 250 terms), all noun terms and place names. Our line chart implemented by icharjs [17], it is a graphics library based on HTML5.

Users can gain the knowledge in two ways. One way is to find interesting terms by a word cloud, then to select one or more interesting terms. The line chart will be useful to illustrate the terms evolution on time series. The other way is to find out the interesting terms from checking the hot terms' line charts, and to check them with word cloud in a particular year.

2.3 Text Cluster and Text Retrieval System

Although many visualization methods intuitively show some features of the text, they cannot be completely replaced by artificial reading. How can we select out the text needed by users and classify them? The text clustering and text retrieval system can fulfill the needs. Users can select the interesting terms as the feathers. Based on these features the text is clustered into several groups based on these features to classify the text. The full-text retrieval assists users to locate interesting words and select the text.

We use the K-means algorithm to do document clustering based on the result of similarity with angle cosine.

$$\mathrm{Sim}\left(d_i, d_j\right) = \frac{\sum_{k=1}^{n} \omega_{ik} \cdot \omega_{jk}}{\sqrt{(\sum_{k=1}^{n} \omega_{ik}^2)(\sum_{k=1}^{n} \omega_{jk}^2)}}, i = 1, 2, \ldots, \mathrm{j} = 1, 2, \ldots \tag{3}$$

Where ω_{ik} and ω_{jk} are the term weighting of term I and term j in document k, respectively. The model of full-text retrieval uses Lucene, which is a highly scalable and super-fast open-source search engine. It offers a simple but powerful program interface. In this model, user can search the text both by the initial input term and by the splitted terms. All of the results will be highlight to help users to see clearly.

3 Case Study

China's foreign trade development is an important part of Chinese culture. Maritime transport is the most important mode of transportation. Therefore, modern coasting trade can reflect China's foreign trade in miniature. After the order of opening ports in 1685, the Qing government set up four customs. They were Canton Customs, Foochow Customs, Chekiang Customs and Shanghai Customs. Compared with the other customs, canton customs existed long and kept prosperous. We now compare our system with N-Gram in the background of Chinese culture in terms of Qing customs.

3.1 Visualization of Customs in the Qing Dynasty Based on N-Gram

N-Gram is the most common tool in cultruomics, in support of many languages such as English, Chinese, and etc. So we try to use it for mining the text about canton customs. Chinese text is still very limited in the Google's library, So we do the following experiments using English and Chinese, respectively.

Now assume we want to inspect the evolution about Canton Customs (粤海关), Foochow Customs (闽海关), Chekiang Customs (浙海关) and Shanghai Customs (江海关). First of all, we use N-Gram to query by the names of these customs. Because the tool is case sensitive, the final result is to sum the number of the result by the customs name of different cases. The key words of query are the English name of these four customs, and we query the words frequency in the period of 1800–2008. Figure 2 shows that before 1842, only canton customs have words frequency for Canton Customs was the most important customs in the early Qing dynasty. As you can see in Fig. 2, after 1842 all of words frequency of these customs became significant. We know that following the Opium War of 1840 China declined to a semi-feudal and semi-colonial country, and the four customs began to trade.

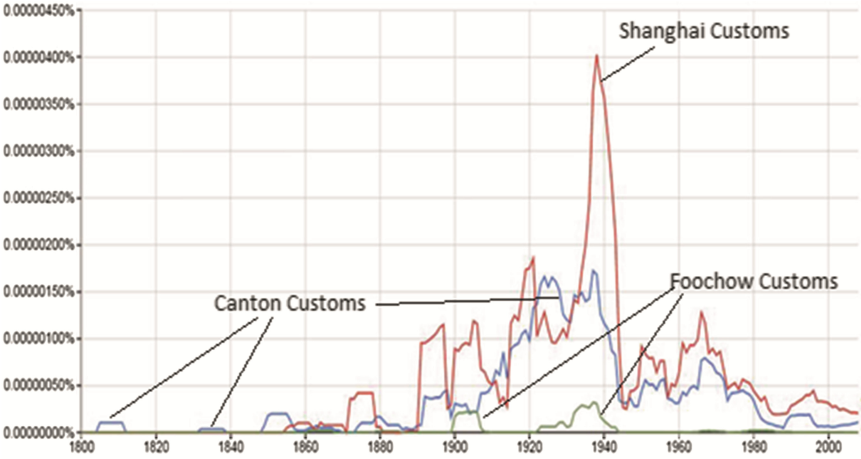

Fig. 2. English words frequency of four customs in Qing Dynasty (Color figure online)

We try to query the words frequency from 1800 to 2008. The Chinese names of these customs are the key words, which are "粤海关, 江海关, 浙海关, 闽海关". As shown in the Fig. 3, there is not any words frequency before 1921 and no any word frequency of Chekiang Customs from 1800 to 2008. It shows that the contemporary Chinese text is relatively scarce in Google Book Library. There is hardly any Chinese text about these four customs during the Qing dynasty. Because of the limitation of Google Book library, we can't study the canton customs based on Chinese text through N-Gram.

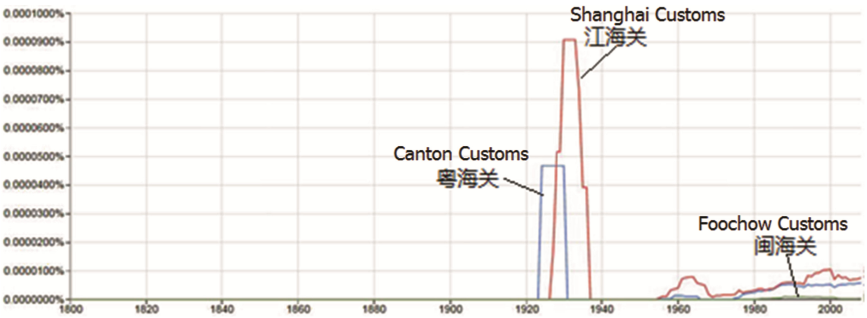

Fig. 3. Words frequency of four customs in modern times (Color figure online)

N-Gram provides only one visual way line chat. The single form limits the reflection of the trend of words frequency. Users need more diversified visual ways to observe the text. In addition, N-Gram is a tool based on the Google Book Library, which hinders users to select and supplement particular contents. We cannot use N-Gram to research the historical text of canton customs. Therefore, A Chinese text mining tool based on Culturomics is required.

3.2 Canton Customs

Canton customs has been an important foreign trade port. Especially, it was the only foreign trade port for over 80 years from 1757 to 1842. At that time canton customs was synonymous with china customs or Qing customs [18].

Our system can have a free supplement or choice. By scanning and electronization, we attain the new text from "A summary of reports on Economic & Social Development Situation in morden GuangZhou Port: The proceeding of canton customs" [19]. The book contains 80 trade reports from 1864 to 1940 and 5 decade reports from 1882 to 1931. The electronized documents are stored in word document format.

To help researchers discover knowledge, in the following research we will use our system on canton customs text in two ways.

Analysis Result of Traded Goods. Figure 4(a) and (b) respectively illustrate us the relatively active goods in 1882 and 1890 silk, including real silk, raw silk and native silk and so on, tea and cinnamon in the forms of the noun word clouds. It is easy to find them in the figures. Then, we chose these terms. Through the line chart we observe them changes from 1870 to 1892 (Fig. 5). We find the raw silk, tea are of great importance in canton customs. Silk was mainly exported as the raw material. On the contrary the occurrence of the cinnamon is lower.

In order to prove the credibility of word frequency line chart, we extracted and counted up the real port data of tea by year. The word frequency of tea is visualized by Fig. 6(a) and the data of real tea export is visualized by Fig. 6(b). We find they have the same varying trend which shows terms frequency can get a reference value results.

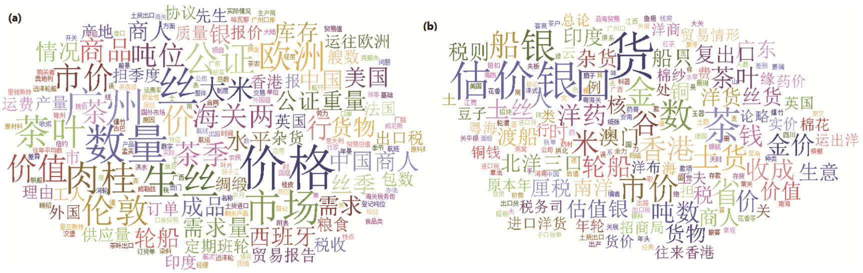

Fig. 4. Chinese words frequency in Qing dynasty

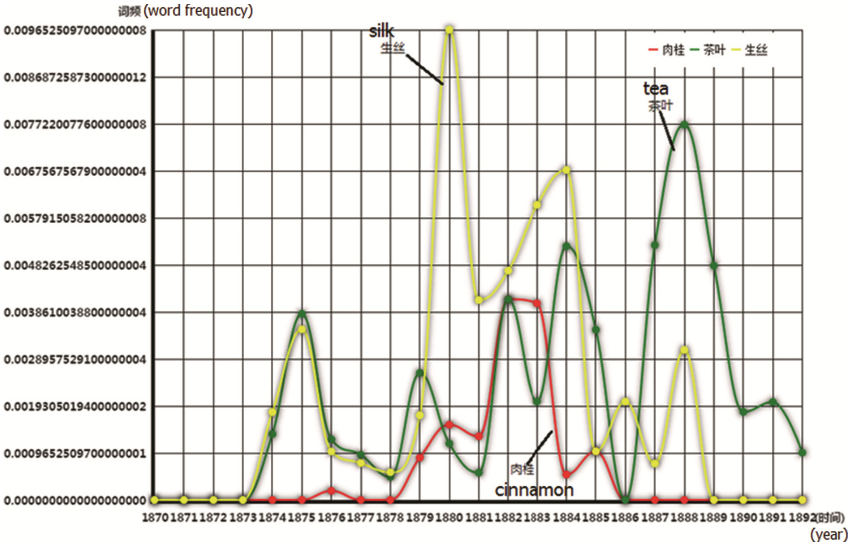

Fig. 5. “生丝” word frequency link chart from1870 to 1892 (Color figure online)

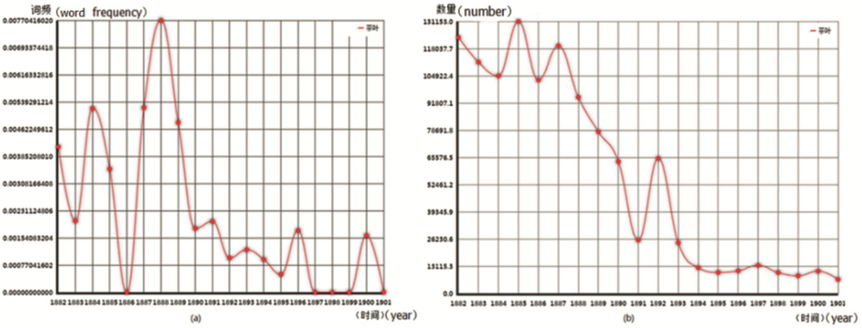

Fig. 6. Comparison chart of tea words frequency and exports data

Analysis Result of Overseas Places. Trade zone is also very important in the study of foreign trade. Our system automatically extracts the highest frequency in five overseas places, including America, Japan, India, Europe and Britain, based on the existing text and displays the results in Fig. 7. Except for Japan, other countries' words frequency fluctuates in the smooth. Before 1894 the word frequency of Japan is almost zero. While after 1894 its frequency is relatively higher. In the second half of the 19th century Japanese successful transformation on economical, increased its desire for developing broad markets. To some extends, it leaded to the outbreak of Sino-Japanese war which broke out in 1894 and China ended in failure [20]. After the war, Japan actively participated in China's foreign trade and maintained its actively for a few decades.

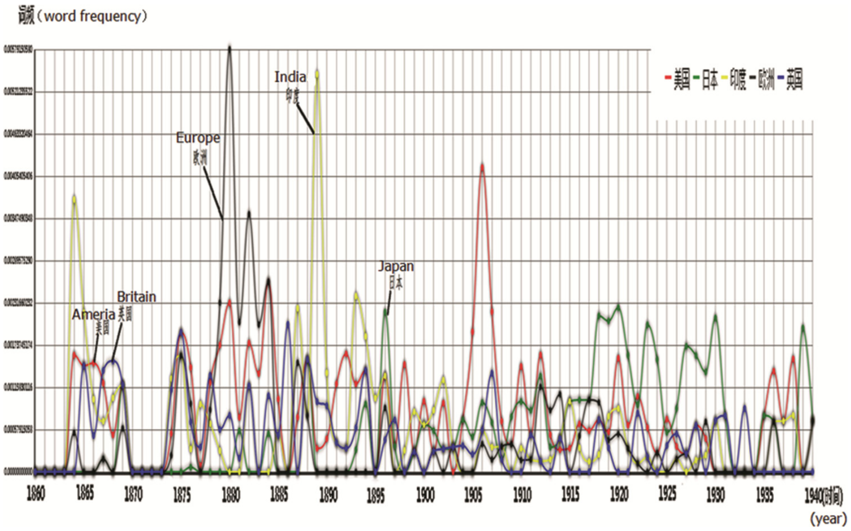

Fig. 7. Five overseas places name of highest words frequency (Color figure online)

Figure 7 shows that India continuous briskness, especially around 1865. Then we view the trade places word clouds (Fig. 8), and find some overseas places which are Havana, Peru, Mexico, were trading with China. In Fig. 9, we observe that their names are only mentioned before 1882 during the eighty years. They were all colonies at that time. We infer that the colonial countries not only plunder the valuable things from these colonies but also use them as trade ports.

Users can select related text using the text clustering and text retrieval system for their intensive reading. Firstly, users need to select the feature words and to input the number of clustering. Then system will provide the results of clustering. In our experiment we chose "Havana, Peru, Mexico" as features and set the clustering in two classes. The detail results is shown as below. The first class includes the trade reports of 1874, 1876 and 1878. The second class includes the rest. Because of the results of word frequency, we can select the first class text. So that researchers can read intensively for their further study.

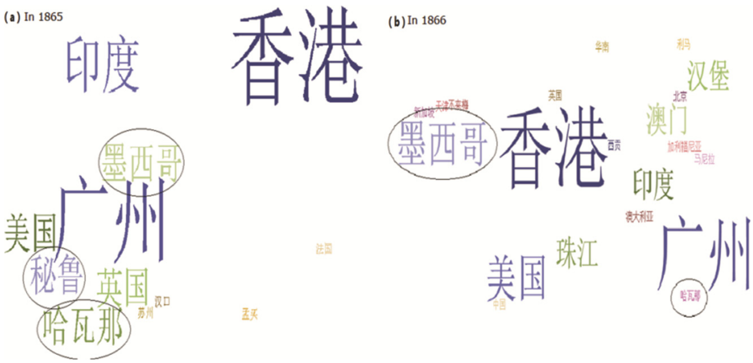

Fig. 8. Word cloud of trade places in 1865 and 1866

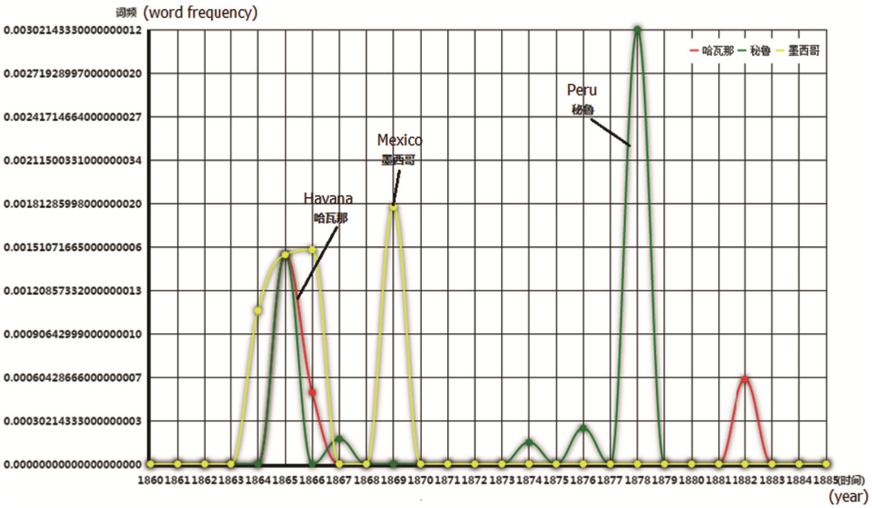

Fig. 9. Word frequency line chart of Havana, Peru, Mexico (Color figure online)

4 Conclusions

In recent years, it has become the trend to analyze and research humanities by quantitative mathematical approaches. In this paper, we established the Chinese historical mining system based on Culturomics. The system includes text data process and analysis subsystem, text data visualization subsystem and also text data collecting and research subsystem. Finally our system provides plentiful visualization results as well as text data clustering and retrievalling to help researcher discover the related information.

Through the case study, compared with N-Gram we have demonstrated the usability of our system. Firstly, we uses N-Gram tool to query the result which reflects the outstanding status of Qing Dynasty Customs. Because of the lack of Chinese resource for Qing Dynasty Customs in the Google database, it is not possible to use N-Gram too for further research. While mining the trade commodity, it can easily spot the frequently traded commodities. While mining oversea places, we found that the colonized area as a trading port by colonial country. Therefore, to establish a Chinese historical text mining system based on Culturomics is a significant exploration.

Our design also has some limitations. We will promote the system function in three ways as below.

- To filer the synonyms to reduce the dimension, labor cost is huge and the result is not comprehensive and accurate. In the next step we will consider to replace it with ontology.
- Based on K-means, text clustering can save time, but the results are not stable due to the initial chosen points. Using a more proper clustering algorithm will give a better result.
- In the visualization, it is not able to observe data efficiently. Using word cloud and line chart make users' operation more complex. Developing a novel method to present the result will comfort researchers to discover information.

Acknowledgements. This work is supported by National Nature science Foundation of China(Grant No. 71171030)

References

1. Shao, P., Lin, Q.: Exploration of extracting chinese cultural genes and modeling its characteristics. J. Xuzhou Normal Univ. (Philos. Soc. Sci. Ed.) **38**, 107–111 (2012)
2. Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., Van Alstyne, M.: Computational social science. Science **323**, 721–723 (2009)
3. Michel, J.B., Shen, Y.K., Aiden, A.P., Veres, A., Gray, M.K., Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M.A., Aiden, E.L.: Quantitative analysis of culture using millions of digitized books. Science **331**, 176–182 (2011)
4. Guo, C., Wei, W., Ren, X.: A review on culturomics. J. China Soc. Scient. Tech. Inform. **33**, 765–774 (2014)
5. Aiden, E.L., Michel, J. (eds.) Uncharted Big Data as a Lens on Human Culture (2013)
6. Chen, Z., Zhang, G.: Study on the text mining and chinese text mining framework. Inform. Sci. **25**, 1046–1051 (2007)
7. Huang, X., Zhao, C.: Application of text mining technology in analysis of net-mediated public sentiment. Inform. Sci. **27**, 94–99 (2009)
8. Gamon, M., Aue, A., Corston-Oliver, S., Ringger, E.: Pulse: mining customer opinions from free text. In: Famili, A., Kok, J.N., Peña, J.M., Siebes, A., Feelders, A. (eds.) IDA 2005. LNCS, vol. 3646, pp. 121–132. Springer, Heidelberg (2005)
9. Wen, T., Yang, D., Li, J.: Design and implementation of Chinese software reviews mining system. Comput. Eng. Des. **34**, 163–167 (2013)
10. He, Q., Zhuang, F.: Big data mining and the Cloud Services. High Technol. Ind. **8**, 56–61 (2013)
11. ICTCLAS:An Open Source Chinese Lexical Analysis System. http://www.nlpir.org
12. Apache. Lucene. http://lucene.apache.org/core

13. Gansner, E.R., Hu, Y.: Stephen: visualizing streaming text data with dynamic graphs and maps. In: 20th International Symposium, Redmond, WA, USA, pp. 439–450 (2012)
14. Cheng, X., Zhu, Q.: Text Mining Principle. Science Press, Beijing (2010)
15. Lee, B., Riche, N.H., Karlson, A.K., Carpendale, S.: SparkClouds: visualizing trends in tag clouds. IEEE Trans. Vis. Comput. Graph. **16**, 1182–1189 (2010)
16. Bostock, M., Ogievetsky, V., Heer, J.: D(3): data-driven documents. IEEE Trans. Vis. Comput. Graph. **17**, 2301–2309 (2011)
17. Ichartjs Open Source Technology Group. http://www.ichartjs.com/
18. Chen, T., Liang, E.: YUEHAIGUANZHI and his researches on the history of the customs. J. Hist. **135**, 72–80 (2009)
19. Office of Local Chronicles Compilation of GuangZhou Province. A summary of reports on Economic & Social Development Situation in morden GuangZhou Port: The proceeding of canton customs. Jinan University Press, GuangZhou (1995)
20. Liu, N.: Is the outcome of the 1894 Sino -Japanese warevitable? -A review from econom ic perspective. Historical Review, pp. 105–115 (2011)